Navigating the Sequence Read Archive to identify crAssphage, an ubiquitous inhabitant of the human microbiome

Jasmine X. Cai[1], Jania G. Weathers[2], Haley Leffler[3], Sruthi Ganapaneni[3], Bhavya Papudeshi[4], Sheri A. Sanders[4], Thomas G. Doak[4]

[1]Center Grove High School, [2]Scecina Memorial High School, [2]Department of Human Biology, Indiana University; [3]National Center for Genome Analysis Support, Pervasive Institute of Technology, Indiana University

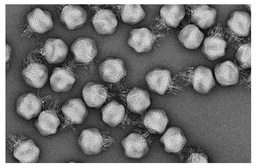
Introduction

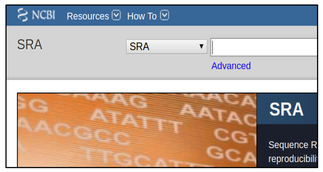
- There are 9 trillion more bacterial cells in the body than human cells (Sender et al., 2016).
- The gut microbiome includes bacteriophages, which are viruses that infect bacteria.
- 73% of humans have crAssphage in their guts, but the role of this phage is unknown (Edwards et al., 2019).
- CrAssphage has been previously found in gut metagenomes, humans, and some water samples. CrAssphage-like genomes have also been identified in primates, but have not been identified in other species.
- In this study we tested to see if crAssphage genome can be found in other species using Sequence Read Archive (SRA).
- SRA hosts 14 PB of genomic data collected all around the world.
- A workflow was previously developed to navigate the SRA and was available on GitHub.
- We tested this workflow, which navigates the SRA to identify datasets that have crAssphage.

Objective

- Our objective is to determine if crAssphage is found in association with other species in the SRA database. This will help us take a step towards determining potential applications for this virus.
- We plan to use a developed workflow and optimize it to help researchers navigate the SRA to efficiently obtain accurate results.

Workflow


Electron micrograph of crAssphage by Rob Edwards via Forbes

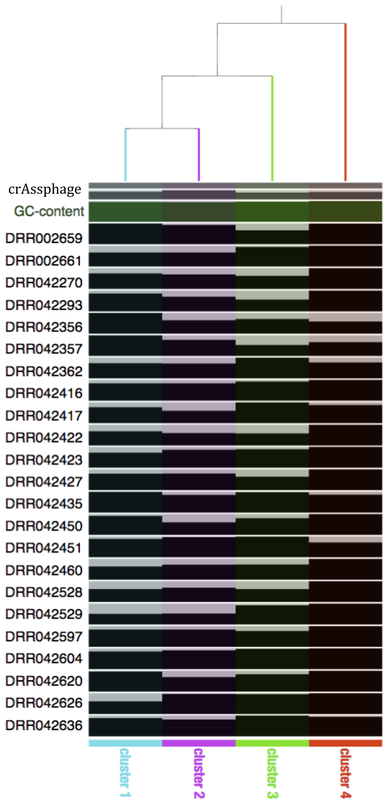


Search Sequence Read Archive (SRA) website hosted by NCBI


This is an image of Jetstream, a cloud computing machine that hosts virtual machines


A cartoon of the researchers going through DNA sequence

Results



Distribution of crAssphage among the twenty three datasets identified from SRA

- The database identified 7500 files that we filtered to predict 23 datasets having crAssphage.
- The information about the datasets from the database suggests that all 23 samples are from human gut.
- We confirmed that crAssphage is found in humans by using Anvi'o, a visualization program to show the 23 files and how they align to the crAssphage.
- These results confirm that crAssphage is only found in humans.
- We optimized the workflow to improve redundant steps, to help researchers look up data in SRA.

Reading the Anvi'o figure:
The figure shows the mean coverage of each cluster within genome samples. The fuller each bar is on the figure, the greater the number of metagenomic reads for that section of the reference genome.

Conclusion

We conclude that crAssphage is found in humans; researchers can now search for characteristics with other bacteriophages that are in humans to identify the function of crAssphage. The workflow applied in this study will also help researchers access the SRA and help them learn more about other organisms.

Big Red II Supercomputer

Jasmine and Jania in front of the Big Red II supercomputer.

- The Summer Science Research Program was a great experience for us, and we gained so much invaluable knowledge just from being here for a week, such as DNA sequencing and bioinformatics.
- We toured the Data Center, where we saw supercomputers. Fun fact - One CPU uses one KW energy which is equal to the energy utilized one house.
- We also did some fun math to find out that in order to store the data in SRA, we would need over 56,000 computers with 250 GB storage.


This is a photo of Jasmine, Jania, and Bhavya on a tour of the Data Center.

Acknowledgements